Fred Dickinson fredsterdickinson@gmail.com

Types of Average: the mean, median, and mode.

Read the notes and attempt all exercises as you go.

Data is useful. There's absolutely no doubt about it. However, it's only useful if we have some way to interpret it. Data sets in the real world are *incredibly* large - try to imagine sitting down in front of a data set with over one million entries! Where would you start?

This is why we have statistics that attempt to capture what the data is telling us. They are trying to *summarise* the data, and are aptly named **summary statistics.** The most well known of these are types of averages, namely the mean, median, and mode. Others include the variance and standard deviation which we will see in another set of notes.

(a) The mean is the 'arithmetic average' of the data. In English terms, 'add up all of the data entries, and then divide by how many entries there are.'
(b) The median is the 'middle' of the data. The data *must be sorted* from low-high for the median to make any sense.
(c) The mode is the 'most frequent' data entry. It is less popular than the mean and median because most realistic data sets have very diverse entries. That being said, it has its moments!

It is important to understand that each type of average has its flaws. To see this, consider the following data set y,

$$y: \quad 32,000 \quad 27,000 \quad 35,000 \quad 176,000 \quad 40,000 \quad 24,000$$

representing the annual salaries (GBP) of six randomly selected individuals.

Exercise 1.1. *What is the mean of the data set y? Do you think this is a good average for the data; does it represent the data well?*

Space for solution. Please turn over when completed.

Solution. The mean of the data set is, to three significant figures, £55, 700. The notation $\overline{y} = 55,700$ is often used to represent the mean. This is *not* a good representation for the average of the data because the vast majority of entries are very far from it.

This data set is small enough for us to spot the problem immediately: the entry of £176, 000 is significantly larger than the rest. We call this value an *outlier* and say that it is *skewing* our mean to be larger than it should be. We can check this by considering the data set *without* the outlier value,

$$y^* : \quad 32,000 \quad 27,000 \quad 35,000 \quad 40,000 \quad 24,000,$$

and then we can calculate $\overline{y^*} = 31,600$. This is a much 'better' average. Notice how dramatically the mean has changed now the outlier entry has been removed.

Unfortunately it is not always realistic to remove outliers, and lots of data sets have entries that skew results. Take, for example, the salaries of workers in the UK. There are a small number of people with *incredibly* high salaries that artificially increase the mean salary.

In these instances the median is often used instead. It is far less susceptible to being skewed by extreme entries[1] whilst remaining representative of the data.

Perhaps now is a good time to restate the most important idea from these notes: *each type of average (and summary statistic in general) has its flaws.* It is so important to recognise this.

Exercise 1.2. *In an advertising survey a small selection of six individuals were asked for their opinions on Marmite. They were told to rank Marmite on a scale from* 1 *to* 10*, where* 1 *indicates 'I strongly dislike this product', and* 10 *indicates 'I love this product.' After sorting, the data looks like*

$$z : \quad 1 \quad 2 \quad 8 \quad 9 \quad 9 \quad 9.$$

Find the mean, median, and mode of this data set. Comment on how well they describe the data. If you had to use one summary statistic, what would you use?

Hint: the 'middle' of a data set with even entries can be found by taking the average of the two middle-most entries.

No model solution is provided. If you'd like feedback on yours, please email me!

[1] See the Office for National Statistics quoting '*[the median is] ... our preferred measure of average earnings, as it is less affected by a relatively small number of very high earners than the mean.*'